La découverte des règles d'association dans un contexte distribué avec des données manquantes : Décomposition tensorielle

Isam El Ayyadi, Mourad Ouziri, Salima Benbernou, Muhammad Younas

La découverte des règles d'association dans un contexte distribué avec des données manquantes : Décomposition tensorielle.

Isam El Ayyadi *, Mourad Ouiziri *, Salima benbernou *,Mohamad Younas **

*Université Paris Descartes, Sorbonne Paris Cité, France
{ isam.el-ayyadi,mourad.ouiziri,salima.benbernou }@parisdescartes.fr,
http://lipade.mi.parisdescartes.fr/
** Oxford Brookes University, Oxford, UK
m.younas@brookes.ac.uk
http://cms.brookes.ac.uk/staff/MYounas/

Résumé. Le recueil des flux de données à travers des capteurs est devenu essentiel dans notre vie de chaque jour, allant de la surveillance du trafic en temps réel aux interventions d'urgence et de surveillance de la santé. Le volume des données entrant est généralement trop élevé pour être stocké et les calculs sur le flux doivent être exécutés en temps réel pour détecter rapidement des événements intéressants (par exemple : la détection et la notification d'accident, le contrôle de congestion du réseau, gestion des pannes de réseau, détection d'intrusion). Cependant, plusieurs de ces événements isolés peuvent également être conjointement surveillés et corrélés afin d'adapter le comportement du système et de prendre les mesures appropriées lors d'une détection d'anomalie.

Dans cet article nous présentons une nouvelle technique permettant de découvrir les règles d'association manquantes dans un réseau multimodal. L'approche proposée est basée sur la décomposition d'un tenseur de confiance avec des valeurs manquantes. Elle est validée par des résultats expérimentaux qui montrent son importance et sa viabilité.

1 Introduction

Aujourd'hui, chaque organisation est confrontée à la manipulation d'une quantité importante de données qui proviennent de sources multiples : données météorologiques, les données des capteurs, les pagesWeb etc.

De nombreux événements intéressants peuvent être détectés par la fouille de ces données provenant de différentes sources distribuées et les analyser à des fins spécifiques Byung-Hoon Park (2002).

Prenant l'exemple de l'industrie automobile Kargupta (2012), les futurs systèmes d'assistance au conducteur devront découvrir, recueillir et analyser des informations dynamiques sur l'environnement de la voiture et de l'état du conducteur. Pour cela Les données seront recueillies à partir de différentes sources qui sont distribués à travers différents endroits.

Cependant la perte d'information et des erreurs dans le processus de collecte sont les deux principaux facteurs qui contribuent à des données manquantes. La conséquence est que certains jeux de données importants peuvent être jetés ou mal analysés produisant des informations incorrectes Kanishka Bhaduri (2011). Les questions mentionnées ci-dessus seront étudiées dans le présent document dans un nouveau type d'environnement distribué à savoir le cloud computing.

Cet article aborde la question de découverte et de prévoir les règles d'association manquantes à partir de données incomplètes sur un nœud de nuage, en les corrélant avec des données provenant d'autres nœuds.

L'approche proposée est basée sur la décomposition tensorielle Tamara G. Kolda (2009). Les décompositions sont appliquées à des tableaux de données pour l'extraction et l'explication de leurs propriétés. Les offres proposées avec un réseau multimodal où les règles d'association manquants sont détectés et leurs confidences sont estimés. Pour cela, les règles d'association à savoir leurs confidences seront représentés sous forme de tableaux dans chaque nœud, où les tableaux obtenus sont incomplets et les résultats de la corrélation entre l'association règles avec d'autres nœuds sont représentés par un tenseur. En d'autres termes, notre objectif à la première tentative est de capturer la structure latente des données via d'ordre supérieur de factorisation en présence de règles d'association. La deuxième tentative est de récupérer les entrées manquantes vers une corrélation distribuée des règles d'association sur le réseau.

Pour valider les résultats obtenus, l'approche distribuée est discuté avec des expériences numériques sur des ensembles de données simulées en présence de données incomplètes et manquantes.

2 État de l'art

Dans le domaine de la fouille des données, l'extraction des règles d'association est une méthode populaire pour découvrir des relations intéressantes entre les variables dans les grandes bases de données Fayyad (1996). Dans cette section, nous donnons quelques définitions utiles traitant des règles d'association et aussi que les tenseurs.

2.1 Itemsets fréquents

— **Définition 1 :** Le support d'un itemset I, noté $Supp(I)$, est égal au nombre d'objets le contenant $Supp(I) \in [1; |G|]$ (appelé aussi support absolu de I).
La fréquence d'un itemset I, notée $Freq(I)$, est égale à $\dfrac{Supp(I)}{|G|}$.

— **Définition 2 :** Itemset fréquent
Un itemset I est dit fréquent si son support, $Supp(I)$, est supérieur ou égal à un seuil minimal d'objets, noté $minsupp$, fixé par l'utilisateur.

— **Proposition :** L'ensemble des itemsets fréquents forme un idéal d'ordre dans $(2M; \subseteq)$ (par rapport à la contrainte de fréquence) :
Tout sous-ensemble d'un itemset fréquent est aussi fréquent, Tout sur-ensemble d'un itemset infréquent est aussi infréquent.

2.2 Règles d'association

— Les mesures les plus utilisées sont le support et la confiance.
— Pour une règle $R : X \implies Y$, la confiance mesure la probabilité qu'une transaction contenant X contienne aussi Y (càd $P(Y|X) = supp(\frac{X \cap Y}{X})$).
— R est donc valide si $Conf(R) \geq minconf$.

2.3 Tenseurs

Un tenseur $\mathcal{X}$ peut être vu comme une généralisation de la notion de vecteur ou de matrice à plusieurs dimensions Herman. et Mechelen (2001) Appellof et Davidson (1981). L'ordre d'un tenseur désigne le nombre de ces dimensions. La figure 2.3 illustre un tenseur d'ordre 3 dont les dimensions successives sont I, J et K. Par analogie avec les matrices, nous pouvons noter un élément de ce tenseur par $x_{i;j;k}$. Un tenseur est un tableau multidimensionnel. L'ordre d'un tenseur représente le nombre de ces dimensions (ou modes).
Les coupes d'un tenseur d'ordre trois sont : les coupes horizontales $\mathbf{X}_{i::}$, les coupes latérales $\mathbf{X}_{:j:}$ et les coupes frontales $\mathbf{X}_{::k}$.
Les fibre d'un tenseur d'ordre 3 sont : les fibres du premier mode $\boldsymbol{x}_{:jk}$, du deuxième mode $\boldsymbol{x}_{i:k}$ et ceux du troisième mode $\boldsymbol{x}_{ij:}$.

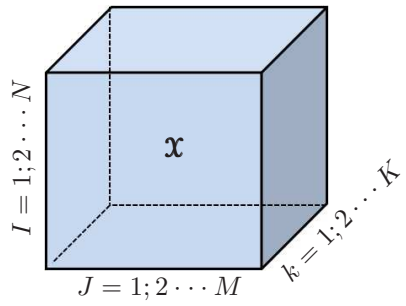

FIG. 1 – *Tenseur* $\mathcal{X} \in \mathbb{R}^{N \times M \times K}$

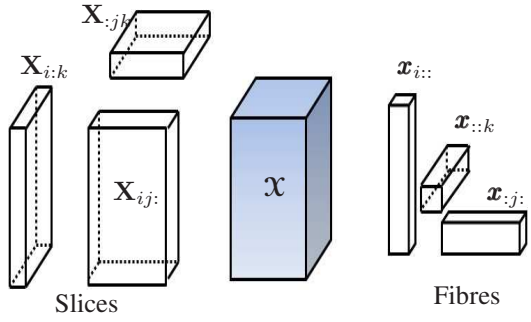

FIG. 2 – *Les coupes d'un tenseur.*

Le produit scalaire et la norme d'un tenseur sont définis par :

$$\mathcal{X} \cdot \mathcal{R} = \sum_{ijk} x_{ijk} x_{ijk}, \qquad \|\mathcal{X}\|^2 = \mathcal{X} \cdot \mathcal{X} = \sum_{ijk} x_{ijk}^2$$

3 Modèle et algorithme de résolution

3.1 Modèle

Dans cette section, on présentera le modèle proposé. L'objectif est de prévoir des règles d'association dans l'environnement de Cloud Computing. dans un réseau multimodal, les données sont réparties entre les différents nœuds $N_1, N_2 ... N_R$ (ou des systèmes informatiques) qui sont reliés par des réseaux. Nous traitons la situation où les données découvertes et extraites

de différents nœuds sont incomplètes.

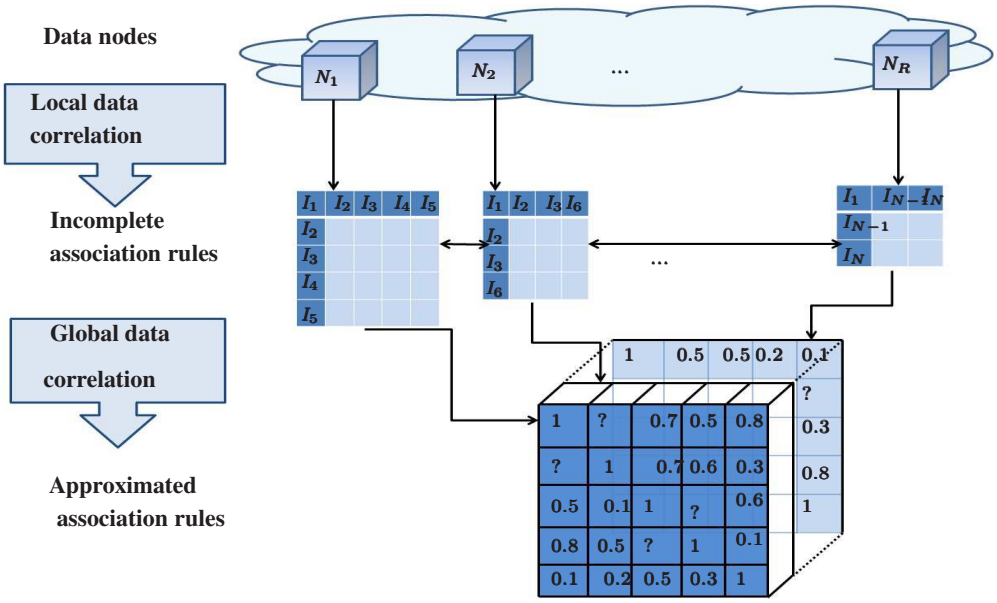

FIG. 3 – *Framework de la corrélation des données distribuées dans le cloud*

3.2 Algorithme

Pour l'extraction locale des relations d'associations, on applique l'algorithme Apriori introduit par Agrawal (1994) qui utilise une méthode bottom-up dans laquelle, à chaque étape, les sous-ensembles fréquents sont élargis d'un item. L'idée de base d'Apriori est qu'un itemset est fréquent si tous ses sous-ensembles sont fréquents.

Le modèle de corrélation 3.2sera représenté par le tenseur $\mathcal{R}$, on notera par ? les valeurs des confiances inconnues. Un fibre $\boldsymbol{r}_{ij:}$ représente les confiances entre les itemsets i et j dans tous les noeuds.

On définit la matrice $\mathcal{W}$ par :

$$\mathcal{W}_{ij} = \begin{cases} 1 & \text{si } X_{i,j,k} \text{ est connu} \\ 0 & \text{si } X_{i,j,k} \text{ est manquant} \end{cases}$$

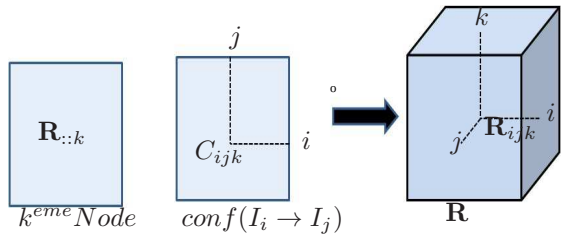

FIG. 4 – *Tenseur de confiance*

Le problème de l'approximation des valeurs manquantes dans le tenseur de confiance $\mathcal{R}$ est équivalent à la recherche d'un tenseur $\mathcal{X}$ qui minimise la forme quadratique suivante :

$$\mathtt{f}(\mathcal{X}) = \sum_{i,j=1}^{N} \mathcal{W}_{ij} \|\mathcal{R}_{ij:} - \mathcal{X}_{ij:}\|^2$$

On applique la décomposition ParaCand decomposition (**CP**)au tenseur $\mathcal{X}$: La forme L de-

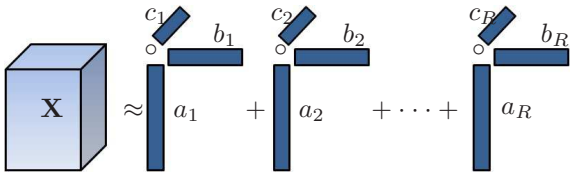

FIG. 5 – *CP*

vient :

$$\mathtt{L}(\mathcal{X}) = \sum_{i,j=1}^{N} \sum_{r=1}^{R} \mathcal{W}_{ij} (\mathcal{R}_{IJR} - \sum_{k=1}^{K} A_{ik} B_{jk} C_{rk})^2$$

Nous appliquerons la méthode du gradient pour minimiser L. Les dérivées partielles de L suivant les trois directions sont données par :

$$\begin{aligned} \frac{\partial \mathtt{L}}{\partial X_{ik}} = & \sum_{j=1}^{N} \sum_{r=1}^{R} \mathcal{W}_{ij} (\sum_{k=1}^{K} X_{ik} Y_{jk} Z_{rk}) Y_{jk} Z_{rk} \\ & - \sum_{j=1}^{N} \sum_{r=1}^{R} \mathcal{W}_{ij} \mathcal{R}_{i,j,r} Y_{jk} Z_{rk} \end{aligned}$$

La découverte des règles d'association : Décomposition tensorielle.

$$\frac{\partial \mathtt{L}}{\partial Y_{jk}} = \sum_{i=1}^{N}\sum_{r=1}^{R} \mathcal{W}_{ij}(\sum_{k=1}^{K} X_{ik}Y_{jk}Z_{rk})X_{ik}Z_{rk} - \sum_{i=1}^{N}\sum_{r=1}^{R} \mathcal{W}_{ij}\mathfrak{R}_{i,j,r}X_{ik}Z_{rk}$$

$$\frac{\partial \mathtt{L}}{\partial Z_{rk}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathcal{W}_{ij}(\sum_{k=1}^{K} X_{ik}Y_{jk}Z_{rk})X_{ik}Y_{jk} - \sum_{i=1}^{N}\sum_{j=1}^{N} \mathcal{W}_{ij}\mathfrak{R}_{i,j,r}X_{ik}Y_{jk}$$

4 Résultats

Dans cette section, nous simulons des données afin d'évaluer la performance de l'algorithme proposé en termes de sa capacité à trouver les valeurs de confiance manquantes.
Pour cela, nous générons une base de données D. Pour chaque nœud, on enlève une partie aléatoire de la base de données. Les bases de données qui en résultent sont : $\{D_1 \cdots D_{10}\}$. L'algorithme Apriori est appliqué à chaque nœud $\{N_i | i = 1 \cdots 10\}$.Le tableau **??** donne des détails sur les statistiques de notre base de données.

TAB. 1 – *Application des statistiques*

Nœuds	10
Nombre total d'Itemsets fréquents	55
Nombre total de confiances manquantes	2360
Tenseur de confiance $\mathcal{X} \in \mathbb{R}^{55\times55\times10}$	30250

Aprés l'application de notre algorithme, nous avons obtenu les résultats suivants :

TAB. 2 – *Model Results*

	N_1	N_2	N_3	N_4
Nombre des Itemsets fréquents	46	50	52	54
Erreur relative ≈	6%	7%	3%	5%

N_5	N_6	N_7	N_8	N_9	N_{10}
45	48	53	48	54	51
6%	7%	8%	7%	1.5%	5%

Globalement, les approximations pour les nœuds 3 et 9 sont intéressants car leurs données sont fermées. Toutefois, les données de nœud et les longerons 7 sont loin de celles des autres nouds, pour laquelle l'erreur d'approximation est la plus élevée.

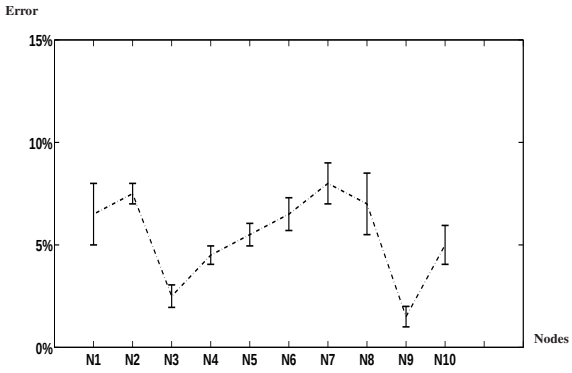

FIG. 6 – *L'erreur relative*

La figure 8 montre l'évolution de l'erreur moyenne en fonction de la quantité de données manquantes. La courbe obtenue montre une certaine stabilité dans notre modèle. En fait, l'effet de l'évolution est linéaire, avec une valeur maximale qui ne dépasse pas 15%.

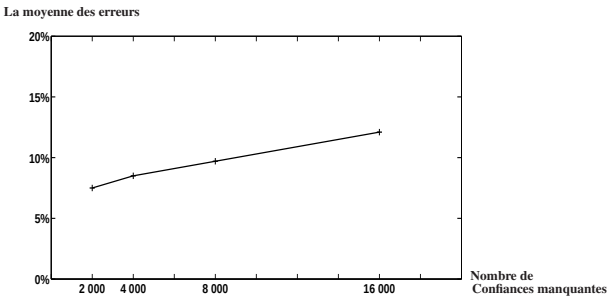

FIG. 7 – *L'évolution de l'erreur*

5 Conclusion

Dans cet article, on a abordé le problème de la découverte des règles d'association manquantes dans le cas où les données sont réparties entre différents nœuds du cloud et certaines données sont manquantes ou erronées. L'algoritme d'approximation est basé sur la décomposition tensorielle. Diverses expériences ont été menées et les résultats obtenus sont convaincants.

Dans l'approche actuelle, le tenseur de confiance considère que les données manquantes ayant des valeurs nulles est basé sur des hypothèses relativement simple. L'avenir comprend l'élargissement du cadre pour gérer (1) hypothèse plus complexe (2) découvrir et analyser les données d'applications en streaming.

Références

Agrawal, R. Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 487–499.

Appellof, C. J. et E. R. Davidson (1981). Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluent. *Anal. Chem*, 2053âĂŞ2056.

Byung-Hoon Park, H. K. (2002). Distributed data mining: Algorithms, systems, and applications. pp. 341–358.

Fayyad, U.M., P.-S. G. S. (1996). From data mining to knowledge discovery : An overview. *An Advances in Knowledge Discovery and Data Mining*, 1–34.

Herman., K. et I. V. Mechelen (2001). Three-way component analysis : Principles and illustrative application. *Psychological Methods 6*, 84âĂŞ110.

Kanishka Bhaduri, Kamalika Das, K. D. B. (2011). Scalable, asynchronous, distributed eigen monitoring of astronomy data streams. *Statistical Analysis and Data Mining 4*, 336–352.

Kargupta, H. (2012). Connected cars: How distributed data mining is changing the next generation of vehicle telematics products. *S-CUBE*, 73–74.

Tamara G. Kolda, B. W. B. (2009). Tensor decompositions and applications. *SIAM Review 51*, 455–500.

Summary

An increasing number of data applications such as monitoring weather data, data streaming, data web logs, and cloud data, are going online and are playing vital in our every-day life. The underlying data of such applications change very frequently, especially in the cloud environment. Many interesting events can be detected by discovering such data from different distributed sources and analyzing it for specific purposes (e.g., car accident detection or market analysis). However, several isolated events could be erroneous due to the fact that important data sets are either discarded or improperly analysed as they contain missing data. Such events therefore need to be monitored globally and be detected jointly in order to understand their patterns and correlated relationships. In the context of current cloud computing infrastructure, no solutions exist for enabling the correlations between multi-source events in the presence of missing data. This paper addresses the problem of capturing the underlying latent structure of the data with missing entries based on association rules.This necessitate to factorize the data set with missing data.

The paper proposes a novel model to handle high amount of data in cloud environment. It is a model of aggregated data that are confidences of associations rules. We first propose a method to discover the association rules locally on each node of a cloud in presence of missing rules. Afterward, we provide a tensor based model to perform a global correlation between all the local models of each node of the network.

The proposed approach based on tensor decomposition, deals with a multi modal network where missing association rules are detected and their confidences are approximated. The approach is scalable in terms of factorizing multi-way arrays (i.e. tensor) in the presence of

missing association rules. It is validated through experimental results which show its significance and viability in terms of detecting missing rules.